Predictions as statements and decisions

Vladimir Vovk

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, England

vovk@cs.rhul.ac.uk

Pittsburgh, PA, June 24, 2006

Major concern of learning theory: prediction (of the label y_n of an object x_n, given a training set).

Standard assumptions in various strands of learning theory:

- statistical learning theory: pairs (x_n, y_n) are independent and identically distributed;
- inductive inference: y_n are generated by a recursive function;
- competitive on-line prediction: no assumptions.

This talk: competitive on-line prediction ($\approx$ subarea of prediction with expert advice).

Different kinds of prediction

- D-predictions: mere Decisions. Never true or false but can be good or bad. Their quality is typically evaluated with a loss function. (Learning theory)
- S-predictions: Statements about reality. They can be tested and, if found inadequate, rejected as false. (Foundations of probability)
- F-predictions (Frequentist predictions): intermediate between D-predictions and S-predictions. Successful if they match the frequencies of various observed events.

This talk: I am really interested in D-prediction. The other kinds: technical tools.

$$\boxed{\text{S-prediction}} \longrightarrow \boxed{\text{F-prediction}} \longrightarrow \boxed{\text{D-prediction}}$$

I will start from a result about D-predictions and then explain how the other kinds can help prove it.

My plan (might be overoptimistic):

- D-prediction: universal consistency
- S-predictions and their testing
- F-predictions: automatic calibration and resolution
- Their use in D-prediction
- Non-asymptotic results about F- and D-prediction
- Conclusion: limitations and questions

Prediction protocol:

FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Predictor announces $\gamma_n \in \Gamma$.
  Reality announces $y_n \in \mathbf{Y}$.
END FOR.

x_n: object (the data relevant for predicting y_n, perhaps including n and some of the previous $y_{n-1}, y_{n-2}, \ldots$;
y_n: label (or observation).

There is a loss function $\lambda(x, \gamma, y)$ (often does not depend on x).

Representative example: $\mathbf{X} = \mathbb{R}^K$, $\Gamma = \mathbf{Y} = \mathbb{R}^L$,
$\lambda(x, \gamma, y) = \|y - \gamma\|$.

Predictor competes with prediction rules $D : \mathbf{X} \to \Gamma$.
Predictor's goal:

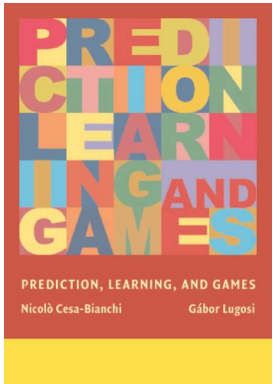
$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) \lessapprox \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n)$$

for all $N = 1, 2, \ldots$ and as many prediction rules D as possible.

Predictor's strategies in the prediction protocol: on-line prediction algorithms.

Sponsored page I

There is a recent excellent review of competitive on-line prediction:



Nicolò Cesa-Bianchi and Gábor Lugosi
Prediction, learning, and games
New York: Cambridge University Press, 2006

Universal consistency for deterministic prediction algorithms

A set in a topological space is precompact if its closure is compact. In Euclidean spaces: precompactness = boundedness.

An on-line prediction algorithm is universally consistent for a loss function λ if its predictions γ_n always satisfy

$$\left(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\right)$$
$$\Longrightarrow \limsup_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) - \frac{1}{N} \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) \right) \leq 0$$

for any continuous prediction rule $D : \mathbf{X} \to \Gamma$.

The loss function λ is large at infinity if for all $x^* \in \mathbf{X}$ and $y^* \in \mathbf{Y}$

$$\lim_{\substack{x \to x^* \\ \gamma \to \infty \\ y \to y^*}} \lambda(x, \gamma, y) = \infty$$

(i.e., for each constant M there exist neighborhoods $A \ni x^*$ and $B \ni y^*$ and compact $C \subseteq \Gamma$ such that $\lambda(A, \Gamma \setminus C, B) \subseteq (M, \infty)$).

Intuitively: faraway γ are bad predictions for nearby y.

Theorem 1a Suppose $\mathbf{X}$ and $\mathbf{Y}$ are locally compact metric spaces, Γ is a convex subset of a Fréchet space, and the loss function $\lambda(x, \gamma, y)$ is continuous, large at infinity, and convex in the variable $\gamma \in \Gamma$. There exists a universally consistent on-line prediction algorithm.

Universal consistency for randomized prediction algorithms

When the loss function $\lambda(x, \gamma, y)$ is not convex in γ: the conclusion of Theorem 1 becomes false (Kalnishkan and Vyugin).

We have to consider randomized prediction rules and randomized prediction algorithms.

A randomized prediction rule is $D : \mathbf{X} \to \mathcal{P}(\Gamma)$. A randomized on-line prediction algorithm is an on-line prediction algorithm in the extended prediction game with the prediction space $\mathcal{P}(\Gamma)$. It is universally consistent if, for any continuous randomized prediction rule $D : \mathbf{X} \to \mathcal{P}(\Gamma)$,

$$\left(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\right)$$
$$\Longrightarrow \left(\limsup_{N\to\infty} \left(\frac{1}{N}\sum_{n=1}^{N} \lambda(x_n, g_n, y_n) - \frac{1}{N}\sum_{n=1}^{N} \lambda(x_n, d_n, y_n)\right) \le 0 \text{ a.s.}\right)$$

where $g_1, g_2, \ldots, d_1, d_2, \ldots$ are independent random variables with g_n distributed as γ_n and d_n distributed as $D(x_n)$, $n = 1, 2, \ldots$.

Theorem 1b Suppose $\mathbf{X}$ and $\mathbf{Y}$ are locally compact metric spaces, Γ is a metric space, and the loss function λ is continuous and large at infinity. There exists a universally consistent randomized on-line prediction algorithm.

Theorems 1a and 1b provide a (maybe partial) answer to Nicolò Cesa-Bianchi's question.

Comparison with the statistical definition I

Statistical setting: the (x_n, y_n) are generated independently from the same distribution. The definition of universal consistency (Stone 1977): similar but no requirement that D should be continuous.

If the requirement of continuity of D is dropped from our definition, universal consistency becomes impossible to achieve for typical loss functions (Reality can choose $x_n \to c$, where c is a point of discontinuity of D).

A positive argument in favor of the requirement of continuity of D: it is natural for Predictor to compete only with computable prediction rules, and continuity is often regarded as a necessary condition for computability (Brouwer's "continuity principle").

Comparison with the statistical definition II

Compactness: also essential in our framework.

Statistical setting: very restrictive. Compactness and continuity may be said to be satisfied automatically: every measurable prediction rule can be arbitrarily well approximated by a continuous one (Luzin's theorem), and every probability measure is almost concentrated on a compact set (Ulam's theorem).

My plan:

- D-prediction: universal consistency
- S-predictions and their testing
- F-predictions: automatic calibration and resolution
- Their use in D-prediction
- Non-asymptotic results about F- and D-prediction
- Conclusion: limitations and questions

Testing predictions

$P \in \mathcal{P}(\Omega)$: a probability forecast for $\omega \in \Omega$. Is it any good in view of the observed ω?

Standard statistical approach: before the experiment, choose $A \subseteq \Omega$ with small $P(A)$ and reject P if $\omega \in A$ (Fisher's disjunction: either P is wrong or a rare event has occurred).

Levin's modification

Levin 1976:
$t : \Omega \times \mathcal{P}(\Omega) \to [0, \infty]$ is a test of randomness if it is lower semicontinuous and, for all $P \in \mathcal{P}(\Omega)$,

$$\int_\Omega t(\omega, P) P(\mathrm{d}\omega) \le 1.$$

If $t(\omega, P)$ is very large for the observed ω, we are entitled to reject P (the P-probability that $t(\omega, P) \ge C$ does not exceed $1/C$).

(In fact Levin required lower semicomputability = efficient version of lower semicontinuity.)

Levin's lemma

Lemma (Levin 1976, somewhat informally in the general case)
Let Ω be a metric compact. For any test of randomness t there exists $P \in \mathcal{P}(\Omega)$ such that

$$\forall \omega \in \Omega : \quad t(\omega, P) \leq 1.$$

Proof: Ky Fan's minimax theorem.

For any test of randomness t there is a probability forecast such that t never detects any disagreement between P and the outcome ω, whatever ω might be.

Digression: universal aspects

Can be applied to the universal test. Gives a neutral probability measure P, such that every ω is random with respect to P.

Gacs (2005): very natural definition of universal test (= largest, to within a constant factor, lower semicomputable test of randomness). Shows that under his definition there are no neutral measures that are computable even in the weak sense of upper or lower computability even in the case of the compactified set of natural numbers.

Levin's original definition of a universal test: extra conditions; Levin's *a priori* measure becomes neutral.

There are no computable neutral measures.

Game-theoretic probability

First seriously explored by von Mises (1919, 1928); important development (martingales): Ville (1939). Defeated by Kolmogorov's measure-theoretic probability (defeat completed by Doob's appropriation of martingales for measure-theoretic probability). Resurrected by Kolmogorov (1963) and Schnorr (1971).

Skeptic = tester

Testing protocol

FOR $n = 1, 2, \ldots$:

Reality announces $x_n \in \mathbf{X}$.

Forecaster announces $P_n \in \mathcal{P}(\mathbf{Y})$.

Skeptic announces $f_n : \mathbf{Y} \to \mathbb{R}$ such that $\int_{\mathbf{Y}} f_n \, \mathrm{d}P_n \le 0$.

Reality announces $y_n \in \mathbf{Y}$.

$\mathcal{K}_n := \mathcal{K}_{n-1} + f_n(y_n)$.

END FOR.

Skeptic's move: a long position in a security that pays $f_n(y_n)$.

$\int_{\mathbf{Y}} f_n \, \mathrm{d}P_n \le 0$ means: $\int_{\mathbf{Y}} f_n \, \mathrm{d}P_n$ exists and is non-positive.

Initial capital: $\mathcal{K}_0$ (not in the protocol).

Testing

Game-theoretic procedure of testing = strategy for Skeptic in the testing protocol.

If Skeptic starts from $\mathcal{K}_0 := 1$, plays so that he never risks bankruptcy, and ends up with a very large value $\mathcal{K}_N$ of his capital, he proves Forecaster wrong.

The role of Skeptic: to detect disagreement between the forecasts and the actual observations. The size of his capital: how successful he is at achieving this goal.

Testing and laws of probability

Law of probability $\approx$ way of testing probability forecasts.

Skeptic can force some property E of the players' moves $x_1, P_1, y_1, x_2, P_2, y_2, \ldots$ in the testing protocol if he has a strategy guaranteeing that (1) his capital $\mathcal{K}_n$ is always nonnegative, and (2) either E is satisfied or $\lim_{n\to\infty} \mathcal{K}_n = \infty$.

The properties that can be forced by Skeptic are the game-theoretic analogue of the properties that hold with probability one in measure-theoretic probability theory.

Sponsored page II

There are many game-theoretic laws of probability collected in

Glenn Shafer and Vladimir Vovk, Probability and finance: it's only a game, New York: Wiley, 2001

All of them exhibit strategies for Skeptic that make him rich when some property of agreement (including various laws of large numbers, the law of the iterated logarithm, the central limit theorem) between the forecasts and the actual observations is violated.

Calibration-cum-resolution theorem for the true probabilities

Theorem 2a Suppose $\mathbf{X}$ and $\mathbf{Y}$ are locally compact metric spaces. Skeptic can force

$$\left(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\right) \Longrightarrow$$
$$\left(\forall f : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y)\right) = 0\right)$$

in the testing protocol, where f ranges over all continuous functions $f : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathbb{R}$.

Follows from the separability of $C(\Omega)$ for metric compacts Ω.

Measure-theoretic result

Corollary Suppose $\mathcal{F}_n$, $n = 0, 1, \ldots$, is a filtration, $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces, $x_1, x_2, \ldots$ are $\mathcal{F}_{n-1}$-measurable $\mathbf{X}$-valued random elements, $y_1, y_2, \ldots$ are $\mathcal{F}_n$-measurable $\mathbf{Y}$-valued random elements, and $P_n \in \mathcal{P}(\mathbf{Y})$ are regular conditional distributions of y_n given $\mathcal{F}_{n-1}$. Then

$$\forall f : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) = 0$$

with probability one.

Proof (you can also easily prove it directly)

The standard recipe for deriving measure-theoretic results from game-theoretic:

- Skeptic's strategy can be chosen measurable (even continuous).
- This makes his capital process $\mathcal{K}_n$ a martingale with respect to the filtration $(\mathcal{F}_n)$.
- This martingale is nonnegative and tends to infinity where the implication fails; standard results of probability theory imply that the implication holds with probability one.

Calibration

What does

$$\forall f : \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) = 0$$

mean?

Take $f(x, P, y) := g(y) I_{P^*}(P)$, where g is continuous and I_{P^*} is a continuous "soft neighborhood" of a fixed $P^* \in \mathcal{P}(\mathbf{Y})$. Then: if we only look at the n for which $P_n \approx P^*$ (and assume there are many such n), the mean of $g(y_n)$ will be close to $\int g(y) P^*(\mathrm{d}y)$; i.e., the empirical distribution of $y_1, \ldots, y_N$ is close to P_n^*. This is called calibration.

Resolution

Similarly, take $f(x, P, y) := g(y) I_{x^*}(x)$, where $x^* \in \mathbf{X}$. Then: if we only look at the n for which $x_n \approx x^*$ (assuming there are many such n), the mean of $g(y_n)$ will be close to the mean of $\int g(y) P_n(\mathrm{d}y)$. This is resolution.

In fact, we have more: if we only look at the n for which $P_n \approx P^*$ and $x_n \approx x^*$ (assuming there are many such n), the mean of $g(y_n)$ will be close to $\int g(y) P^*(\mathrm{d}y)$. Calibration-cum-resolution.

Defensive forecasting

Levin's Lemma: defensive probability forecast P. Can be extended to game-theoretic probability.

Suppose Skeptic has chosen in advance, and told Forecaster about, his testing strategy.

Defensive forecasting protocol

FOR $n = 1, 2, \ldots$:

Reality announces $x_n \in \mathbf{X}$.

Skeptic announces a lower semicontinuous $F_n : \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathbb{R}$

such that $\int_{\mathbf{Y}} F_n(P, y) P(\mathrm{d}y) \le 0$, $\forall P \in \mathcal{P}(\mathbf{Y})$.

Forecaster announces $P_n \in \mathcal{P}(\mathbf{Y})$.

Reality announces $y_n \in \mathbf{Y}$.

$\mathcal{K}_n := \mathcal{K}_{n-1} + F_n(P_n, y_n)$.

END FOR.

(In fact, we need less: Skeptic's strategy is disclosed in a piecemeal manner.)

Takemura's lemma

Lemma (Takemura) Let $\mathbf{Y}$ be a metric compact. In the defensive forecasting protocol, Forecaster can play in such a way that Skeptic's capital never increases, no matter how he and Reality play.

Proof: apply Levin's lemma in each round.

A simple version (with $\mathbf{Y} = \{0, 1\}$) stated by Akimichi Takemura in his comment on Glenn Shafer's talk at the University of Tokyo (May, 2004).

My plan:

- D-prediction: universal consistency
- S-predictions and their testing
- F-predictions: automatic calibration and resolution
- Their use in D-prediction
- Non-asymptotic results about F- and D-prediction
- Conclusion: limitations and questions

Probability forecasting protocol (more basic than the testing and defensive forecasting protocols)

FOR $n = 1, 2, \ldots$:

Reality announces $x_n \in \mathbf{X}$.

Forecaster announces $P_n \in \mathcal{P}(\mathbf{Y})$.

Reality announces $y_n \in \mathbf{Y}$.

END FOR.

Forecaster's strategy: probability forecasting strategy.

Asymptotic result about automatic calibration and resolution

Using defensive forecasting from Theorem 2a:

Theorem 2b Suppose $\mathbf{X}$ and $\mathbf{Y}$ are locally compact metric spaces. There is a strategy for Forecaster in the probability forecasting protocol that guarantees

$$\left(\{x_1, x_2, \ldots\} \text{ and } \{y_1, y_2, \ldots\} \text{ are precompact}\right)$$
$$\Longrightarrow \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \left(f\left(x_n, P_n, y_n\right) - \int_{\mathbf{Y}} f\left(x_n, P_n, y\right) P_n(\mathrm{d}y) \right) = 0$$

for all continuous functions $f : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y} \to \mathbb{R}$.

Kakade & Foster: no $\mathbf{X}$, finite $\mathbf{Y}$, Lipschitz f.

Verging on paradox

Dawid's (1985) example:

For $\mathbf{Y} = \{0, 1\}$, Reality outputs

$$y_n := \begin{cases} 1 & \text{if } P_n(\{1\}) \le 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Forecaster is grossly miscalibrated.

Explanation: if Theorem 2b's Forecaster plays against Dawid's Reality, P_n converges to the uniform distribution U. Skeptic needs stronger and stronger microscope to detect lack of calibration; since Forecaster knows the resolution of his microscope, he can make $P_n \to U$ fast enough.

Remark 1

The first results like Theorem 2b (Foster and Vohra 1998 and many papers developing it) were completely independent of Levin's idea.

Remark 2

Theorems 2a and 2b suggest that probability theory and competitive on-line prediction might enrich each other:

- laws of probability can be translated into useful probability forecasting strategies via defensive forecasting;
- the needs of competitive on-line prediction can help identify and fill gaps in the existing probability theory.

My plan:

- D-prediction: universal consistency
- S-predictions and their testing
- F-predictions: automatic calibration and resolution
- Their use in D-prediction
- Non-asymptotic results about F- and D-prediction
- Conclusion: limitations and questions

Use in D-prediction

D-prediction algorithms can be built on top of probability forecasting algorithms: D-predictions are found by minimizing the expected loss, with the expectation taken with respect to the probability forecast.

Problem: the possibility that the minimizer of the expected loss will be a discontinuous function (and continuity is essential in defensive forecasting).

Define

$$\lambda(x, \gamma, P) := \int_{\mathbf{Y}} \lambda(x, \gamma, y) P(\mathrm{d}y).$$

Problem of discontinuity

$G : \mathbf{X} \times \mathcal{P}(\mathbf{Y}) \to \Gamma$ is a choice function if

$$\lambda(x, G(x, P), P) = \inf_{\gamma \in \Gamma} \lambda(x, \gamma, P), \quad \forall x \in \mathbf{X}, P \in \mathcal{P}(\mathbf{Y}).$$

There is no continuous choice function even in the absolute loss game (no $\mathbf{X}$, $\Gamma = [0, 1]$, $\mathbf{Y} = \{0, 1\}$, $\lambda(\gamma, y) := |y - \gamma|$). Any choice function G must satisfy

$$G(P) := \begin{cases} 1 & \text{if } P(\{1\}) > 1/2 \\ 0 & \text{if } P(\{1\}) < 1/2, \end{cases}$$

but the case $P(\{1\}) = 1/2$ is a point of bifurcation.

Luckily approximate continuous choice functions will often exist.

Proof of a weak form of Theorem 1

$\mathbf{X}$, Γ, $\mathbf{Y}$: compact metric spaces. Then:

for each $\epsilon > 0$ there exists an ϵ-approximate continuous choice function G, satisfying

$$\lambda(x, G(x, P), P) \leq \inf_{\gamma \in \Gamma} \lambda(x, \gamma, P) + \epsilon, \quad \forall x \in \mathbf{X}, P \in \mathcal{P}(\mathbf{Y}).$$

Replace ≤ 0 by $\leq \epsilon$ for an arbitrarily small constant $\epsilon > 0$.

Predictor chooses his decision by applying the approximate choice function G to x_n and probability forecasts P_n for y_n with good calibration-cum-resolution.

$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) = \sum_{n=1}^{N} \lambda(x_n, G(x_n, P_n), y_n)$$

$$= \sum_{n=1}^{N} \lambda(x_n, G(x_n, P_n), P_n) + \sum_{n=1}^{N} \Big(\lambda(x_n, G(x_n, P_n), y_n) - \lambda(x_n, G(x_n, P_n), P_n) \Big)$$

$$= \sum_{n=1}^{N} \lambda(x_n, G(x_n, P_n), P_n) + o(N) \leq \sum_{n=1}^{N} \lambda(x_n, D(x_n), P_n) + \epsilon N + o(N)$$

$$= \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) - \sum_{n=1}^{N} \Big(\lambda(x_n, D(x_n), y_n) - \lambda(x_n, D(x_n), P_n) \Big) + \epsilon N + o(N)$$

$$= \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \epsilon N + o(N)$$

My plan:

- D-prediction: universal consistency
- S-predictions and their testing
- F-predictions: automatic calibration and resolution
- Their use in D-prediction
- Non-asymptotic results about F- and D-prediction
- Conclusion: limitations and questions

RKHS with finite embedding constant

Non-asymptotic results about calibration and resolution: in terms of "reproducing kernel Hilbert spaces".

Let $\mathcal{F}$ be a Hilbert space of functions on a set Ω. Its embedding constant:

$$\mathbf{c}_{\mathcal{F}} := \sup_{\omega \in \Omega} \sup_{f \in \mathcal{F}: \|f\|_{\mathcal{F}} \leq 1} f(\omega).$$

Suppose $\mathbf{c}_{\mathcal{F}} < \infty$. Then $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS): all evaluation functionals $f \in \mathcal{F} \mapsto f(\omega)$, $\omega \in \Omega$, are bounded.

Calibration-cum-resolution bounds

By the Riesz–Fischer theorem, for each $\omega \in \Omega$ there exists a function $\mathbf{k}_\omega \in \mathcal{F}$ (the representer of ω in $\mathcal{F}$) such that

$$f(\omega) = \langle \mathbf{k}_\omega, f \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}.$$

If Ω is a topological space and the mapping $\omega \mapsto \mathbf{k}_\omega$ is continuous, $\mathcal{F}$ is called a continuous RKHS.

Theorem Let $\mathbf{Y}$ be a metric compact and $\mathcal{F}$ be a continuous RKHS on $\mathbf{X} \times \mathcal{P}(\mathbf{Y}) \times \mathbf{Y}$ with finite embedding constant $\mathbf{c}_{\mathcal{F}}$. There is a probability forecasting strategy that guarantees

$$\left| \sum_{n=1}^{N} \left(f(x_n, P_n, y_n) - \int_{\mathbf{Y}} f(x_n, P_n, y) \, P_n(\mathrm{d}y) \right) \right| \leq 2\mathbf{c}_{\mathcal{F}} \|f\|_{\mathcal{F}} \sqrt{N}$$

for all N and all $f \in \mathcal{F}$.

Continuous universal RKHS

RKHS $\mathcal{F}$ on Ω is universal if Ω is a topological space and for every compact subset A of Ω every continuous function on A can be arbitrarily well approximated in the metric $C(A)$ by functions in $\mathcal{F}$ ($\approx$ Steinwart 2001).

Many continuous universal RKHS are known: Sobolev spaces, RKHS with Gaussian kernels, RKHS with infinite polynomial kernels,. . . .

Using such kernels we can also derive the asymptotic Theorem 2b.

Non-asymptotic version of Theorem 1a

For each prediction rule $D : \mathbf{X} \to \Gamma$, define

$$\lambda_D(x, y) := \lambda(x, D(x), y).$$

Set

$$\mathbf{c}_\lambda := \sup_{x \in \mathbf{X}, \gamma \in \Gamma, y \in \mathbf{Y}} \lambda(x, \gamma, y) - \inf_{x \in \mathbf{X}, \gamma \in \Gamma, y \in \mathbf{Y}} \lambda(x, \gamma, y)$$

(finite if λ is continuous and $\mathbf{X}, \Gamma, \mathbf{Y}$ are compact).

Theorem Suppose $\mathbf{X}$ and $\mathbf{Y}$ are compact metric spaces, Γ is a convex compact subset of a topological vector space and the loss function $\lambda(x, \gamma, y)$ is continuous in (x, γ, y) and convex in $\gamma \in \Gamma$. Let $\mathcal{F}$ be a continuous RKHS on $\mathbf{X} \times \mathbf{Y}$ with finite embedding constant $\mathbf{c}_{\mathcal{F}}$. There is an on-line prediction algorithm that guarantees

$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) \leq \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n) + \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2}\,(\|\lambda_D\|_{\mathcal{F}} + 1)\sqrt{N} + 1$$

for all prediction rules D and all $N = 1, 2, \ldots$.

($\|f\|_{\mathcal{F}} := \infty$ for $f \notin \mathcal{F}$)

Non-convex loss functions

Applying Hoeffding's inequality:

Corollary Suppose $\mathbf{X}, \mathbf{Y}, \Gamma$ are compact metric spaces and the loss function λ is continuous. Let $N \in \{1, 2, \ldots\}$ and $\delta \in (0, 1)$. There is a randomized on-line prediction algorithm achieving

$$\sum_{n=1}^{N} \lambda(x_n, g_n, y_n) \leq \sum_{n=1}^{N} \lambda(x_n, d_n, y_n)$$
$$+ \sqrt{\mathbf{c}_\lambda^2 + 4\mathbf{c}_{\mathcal{F}}^2}\,(\|\lambda_D\|_{\mathcal{F}} + 1)\sqrt{N} + \mathbf{c}_\lambda \sqrt{2 \ln \frac{1}{\delta}} \sqrt{N} + 1$$

with probability at least $1 - \delta$ for any randomized prediction rule $D : \mathbf{X} \to \mathcal{P}(\Gamma)$; g_n and d_n are independent random variables distributed as γ_n and $D(x_n)$, respectively.

Remarks

The last 2 results are non-vacuous only when $\lambda_D \in \mathcal{F}$. If $\mathcal{F}$ is a Sobolev space, this condition follows from $D \in \mathcal{F}$ and the smoothness of λ. Result of this type: Moser 1966.

Applying these results to continuous universal RKHS, one can deduce Theorems 1a and 1b.

My plan:

- D-prediction: universal consistency
- S-predictions and their testing
- F-predictions: automatic calibration and resolution
- Their use in D-prediction
- Non-asymptotic results about F- and D-prediction
- Conclusion: limitations and questions

Limitation I

All results I mentioned: one-step-ahead prediction. In a more general framework the loss function would depend not only on y_n but on other future outcomes as well. There are simple ways of extending our results in this direction: e.g., if the loss function $\lambda = \lambda(x_n, \gamma_n, y_n, y_{n+1})$ depends on both y_n and y_{n+1}, we could run two on-line prediction algorithms with the observation space $\mathbf{Y}^2$, one responsible for choosing γ_n for odd n and the other for even n. However, cleaner and more principled approaches are needed.

Limitation II

D-predictions: decisions made by a small decision maker (i.e., his decisions do not affect Reality's future behavior).
Remember Predictor's goal:

$$\sum_{n=1}^{N} \lambda(x_n, \gamma_n, y_n) \lessapprox \sum_{n=1}^{N} \lambda(x_n, D(x_n), y_n).$$

Predictor's and D's losses are compared on the same sequence $x_1, y_1, x_2, y_2, \ldots$. If Predictor is a big decision maker this is unnatural: presumably, $x_1, y_1, x_2, y_2, \ldots$ resulted from Predictor's decisions γ_n, and D's loss should be evaluated on the sequence $x_1^*, y_1^*, x_2^*, y_2^*, \ldots$ resulting from D's decisions $D(x_n)$.

Ideal is not good enough I

Defensive forecasting produces ideal, in important aspects, probability forecasts. Ideal forecasts are not sufficient in big decision making.

Consider the game where there is no $\mathbf{X}$, $\Gamma = \mathbf{Y} = \{0, 1\}$, and the loss function λ is given by the matrix

	$y = 0$	$y = 1$
$\gamma = 0$	1	2
$\gamma = 1$	2	0

Reality's strategy is $y_n := \gamma_n$, but Predictor's initial theory is that Reality always chooses $y_n = 0$.

Ideal is not good enough II

Predictor's "optimal" strategy based on his initial beliefs is to always choose $\gamma_n = 0$ suffering loss 1 at each step. His initial beliefs are reinforced with every move by Reality. Intuitively it is clear that Predictor's mistake in not choosing $\gamma_n \equiv 1$ is that he was being greedy (concentrated on exploitation and completely neglected exploration). However,

- he acted optimally given his beliefs,
- his beliefs have been repeatedly confirmed by what actually happened.

In big decision making we have to worry about what would have happened if we had acted in a different way.

Lesson for defensive forecasting

The skeptic testing the theory should not do this only by playing the imaginary game with the imaginary capital; he should also venture in the real world. Predictor's theory that Reality always chooses $y_n = 0$ would not survive for more than one round had it been tested (by choosing a sub-optimal, from the point of view of the current theory, decision).

Question I: The data space as a bottleneck

If we set $\mathbf{X} := \sum_{n=0}^{\infty} \mathbf{Y}^n$ and

$$x_n := (y_1, \ldots, y_{n-1}),$$

it becomes impossible to compete even with the simplest prediction rules $D : \mathbf{X} \to \mathbf{Y}$: there needs be no connection between the restrictions of D to $\mathbf{Y}^n$ for different n.

$x_n \in \mathbf{X}$: a necessary bottleneck to restrict the competition.

Can the bottleneck (the local compactness of $\mathbf{X}$) be widened?

Question II: Randomization

When using a prediction strategy based on defensive forecasting, Predictor needs randomization only when there are several very different predictions with approximately the same expected loss with respect to the current probability forecast P_n.

Since P_n are guaranteed to agree with reality, we would not expect that Predictor will often find himself in such a position provided Reality is neutral (rather than an active opponent). Predictor's strategy will be almost deterministic.

How to formalize this intuition?

Proofs

Proofs of all results I mentioned can be found in:

Vladimir Vovk, Predictions as statements and decisions. Technical Report `arXiv:cs.LG/0606093`, `arXiv.org` e-Print archive, June 22, 2006.

Latest version (and these slides): `http://vovk.net`.